Cmput 466 Mini-Project

Nicholas Chee

April 13, 2020

Motivation

The Wisconsin Breast Cancer Dataset [1] consists of 699 samples with 30 features collected from a digitized image of a Fine Needle Aspirate (FNA) in order to diagnose whether or not a patient has cancer. In particular, a tumor may be classified as malignant (and therefore the cells inside the tumor are cancerous), or benign (in which, the tumor is normal.) An early diagnosis of such tumors is greatly important; when detected early, there is a 30% chance of effective treatment, compared to late stage treatment.

In this project, we will examine a few different classification techniques to diagnostically predict whether or not a tumor is malignant or benign. We chose the methods of K-Nearest Neighbour (KNN), Logistic Regression, Support Vector Machines (SVM) to conduct such an analysis.

EDA

We can attempt to visualize the underlying distribution of the sampled breast cancer data, and look for any ways to reduce the dimensions of our features. We can then see if any hyperplane/decision-boundary exists between the two labels. The following plots are shown below:

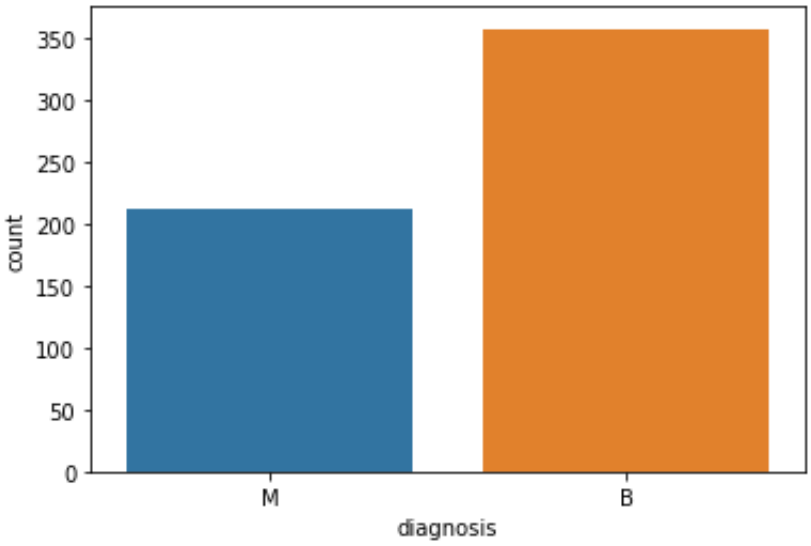

Figure 1: Count plot for visualizing distribution of the data.

Here, we have a larger frequency of benign tumors, although there is still enough malignant samples for different classifiers to learn from. We can also look at performing some form of multidimensional scaling on the dataset in order to see if there exists a clear separation the data. We can first try performing metric distance scaling, which computes the dissimilarities between each feature value and summing them. Note that we normalize our features first so that no higher order scale skews the calculated distances.

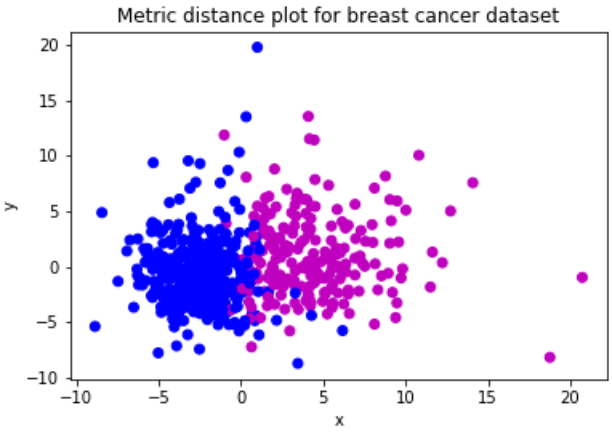

Figure 2: Classical multidimensional scaling plot for the cancer data.

Here, we see that there seems to be a clear separation of the data, where the blue benign samples are on the left of the plot, and the purple malignant samples are on the right. There also exists some overlap between data points, so a hard-margin classifier may perform poorly against a soft-margin classifier.

We can also plot the samples projected onto its principal components via Principal Component Analysis, which finds a projection that maximizes the variance between points.

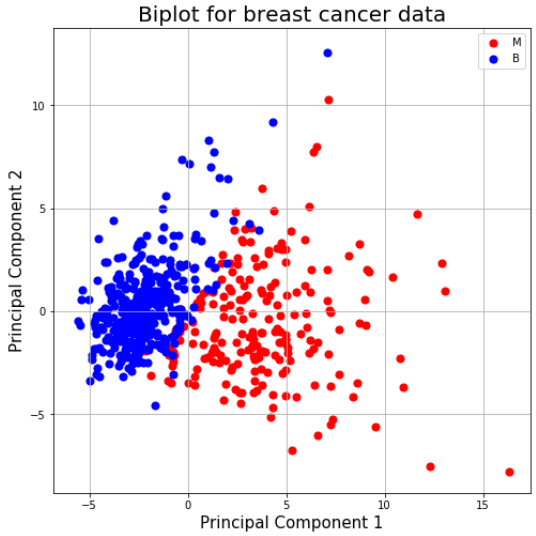

Figure 3: Biplot for the cancer data.

From the above plot, we also get a clear separation of the data. From this exploratory analysis, one might expect classifiers relying on some distance/difference metric between points to perform well in this setting. Thus, all the methods outlined above should be adequate for predicting the existence of cancerous cells.

Model Evaluation

In this section, we look at several classifiers for predicting the diagnosis of tumors in the breast cancer dataset. In particular, we look at *K-Nearest Neighbours*, *Logistic Regression*, and *Support Vector Machines* in order to determine which classification is best for this problem setting. We use the *train-validation-test framework*, splitting the data into 80% labelled, and 20% test. We then further split the labelled data using *10-fold cross-validation* to compute hyperparameter values that yield the best performance among the folds. We use the *F1-measure* to evaluate performance of our classifier, given that we care more about the false negatives since patients with cancer who test negative fail to follow-up on further tests.

Dummy Classifier

A dummy classifier is used as a benchmark to test the other classifier's performance against. For this problem, we chose to use a classifier that generates predictions sampled from a uniform distribution. The results from this kind of classifier fit better than a "majority guess" classifier in that because we use the *F1-measure* as our evaluation metric, the classifier would have a poor performance. We use *scikit-learn*'s dummy classifier method to achieve this, attaining an *F-Measure* of 0.255 on the test set. We also get the following confusion matrix:

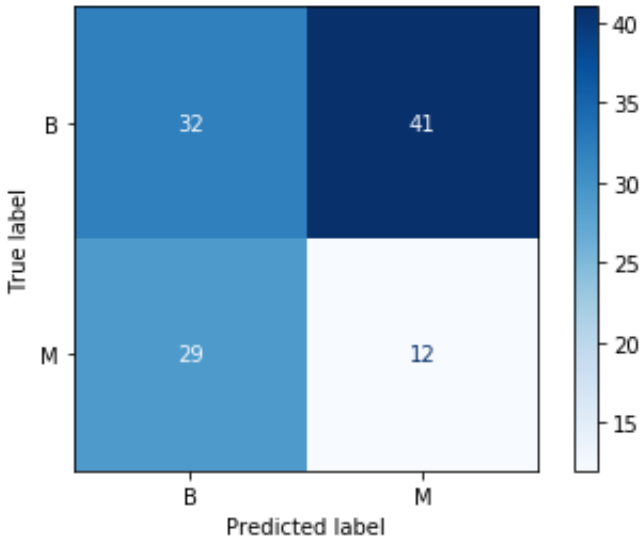

Figure 4: Confusion matrix for the dummy classifier.

Thus, any model that can achieve an accuracy larger than this is substantial.

K-Nearest Neighbours

Using the method of K-Nearest Neighbours, we classify based on the most common class among its k-nearest neighbours. Using the train-validation-test framework, we extracted the best hyperparameter value from 10-fold cross validation using two methods: First, we use grid search with takes the cross-product of all parameters specified, outputting the best combination of such parameters. Then, we use randomized search to sample from the grid.

The grid search method yields a k value equal to 8, and uses distance for the *weights* parameter and euclidean distance as a metric for the distance for each neighbour, whereas the classifier obtained from performing randomized search yields a k value of 3, using uniform weights and Manhattan distance to perform classification. Although its validation performance tested lower at 94.69% versus 95.11%, it achieved a higher *F1-measure* of 98.25% from the test set compared to the 97.37% score obtained from the grid search model. This may be attributed to the lower k value, meaning that the classifier attempts to fit to the training data more.

Below are the learning curves and confusion matrices for the above two methods.

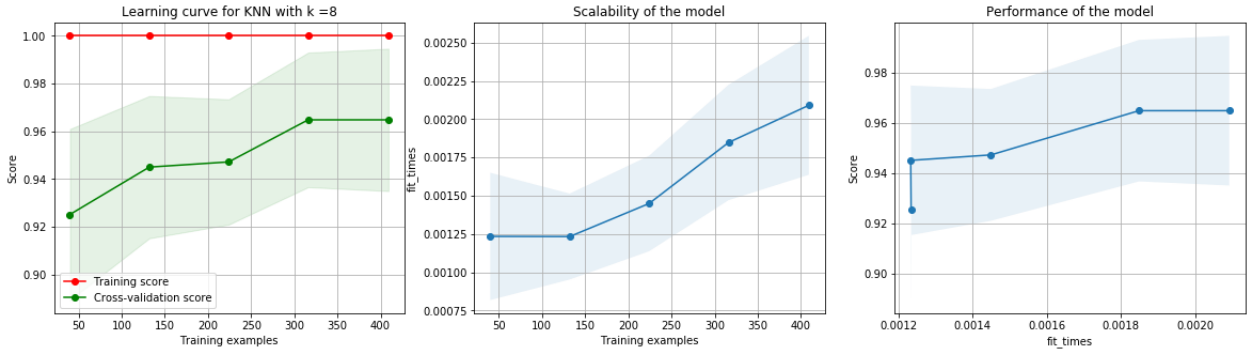

Figure 5: Learning curve, scalability, and performance of KNN model using grid search technique.

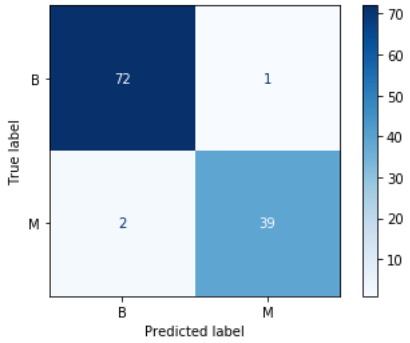

Figure 6: Confusion matrix for KNN model using grid search technique.

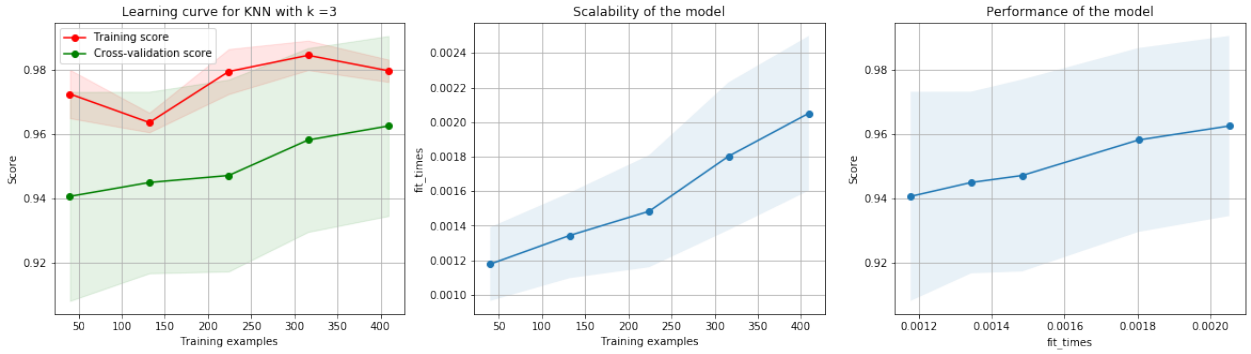

Figure 7: Learning curve, scalability, and performance of KNN model using randomized search technique.

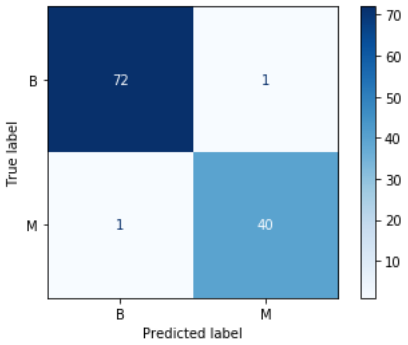

Figure 8: Confusion matrix for KNN model using randomized search technique.

Logistic Regression

In this section, we use logistic regression to classify the diagnosis of types of tumors. Using the same methods as above (i.e. grid search and randomized search), we found that the grid search method yielded a higher *F1-measure* on the test set at 99.12% with hyperparameters $C = 0.695$ with an l_2 penalty. The learning curve and confusion matrix for this classifier are shown below:

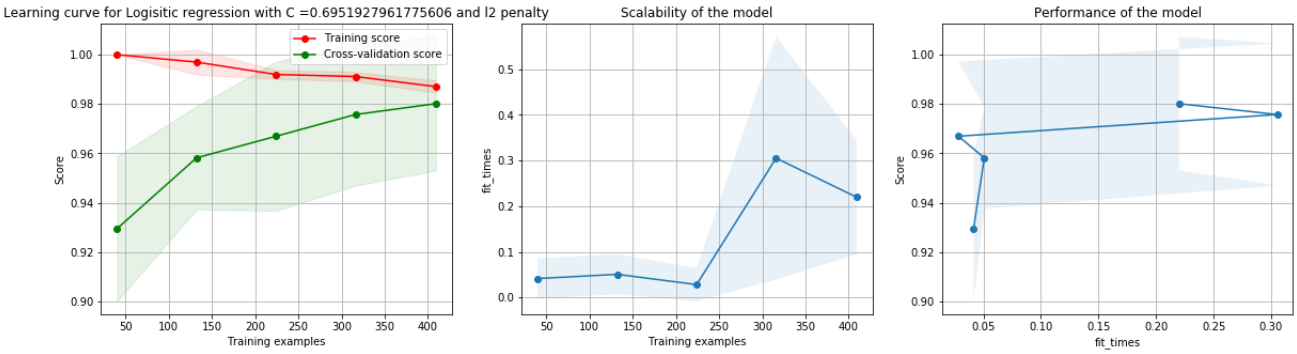

Figure 9: Learning curve, scalability, and performance of logistic regression model using grid search technique.

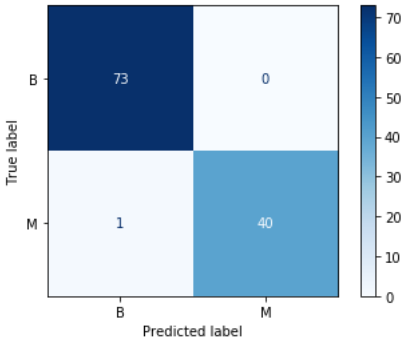

Figure 10: Confusion matrix for logistic regression model using grid search technique.

Because this model yielded a higher test score compared to the KNN method, we can say that the logistic regression classifier is better suited for the problem of classifying types of tumors.

Support Vector Machines

Finally, we use *Support Vector Machines* as a means of classifying whether or not a tumor is malignant, or benign. In the previous section, we note that there exists some overlap between the two classes. Therefore, we can look at using soft-margin SVMs to classify the data. Again, we use the same methods from above, and found that the randomized search method computed a better classifier with a C value equal to 1000 using a radial basis function as its kernel. Against the test set, we computed a *F1-measure* to be 99.12%, with the learning curve and confusion matrix provided below:

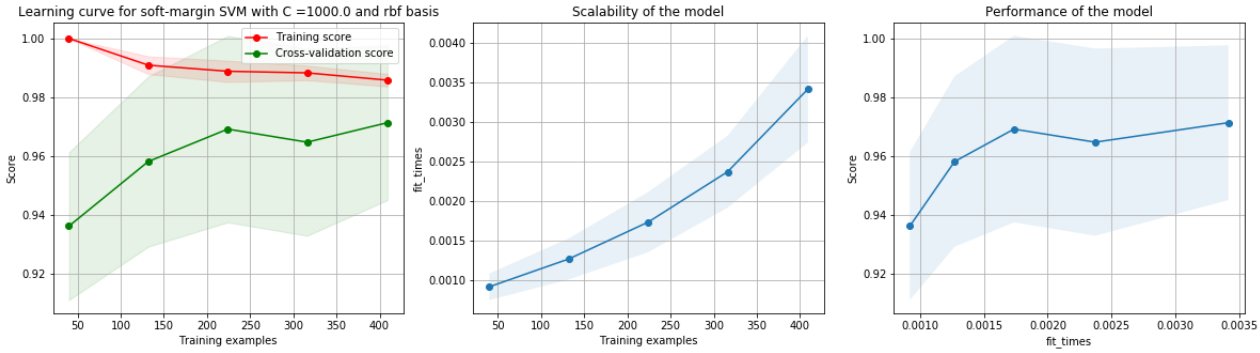

Figure 11: Learning curve, scalability, and performance of support vector machine model using randomized search technique.

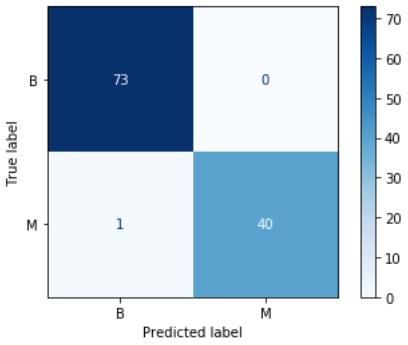

Figure 12: Confusion matrix for support vector machine model using randomized search technique.

Because the scores produced from the SVM model and the logistic regression model are the same, we cannot confidently conclude that one classification method is better than the other.

Conclusion

From our evaluation of the three models, we can conclude that the SVM model and the logistic regression model are both adequate classification models for determining whether or not a tumor is malignant or benign, yielding large *F1* scores compared to the K-Nearest Neighbours method. However, all three models in general produce high scores, especially compared to the dummy classifier. Therefore, each model is substantial and can be used for the breast cancer problem setting given the same features.

References

[1] William H. Wolberg. Breast cancer wisconsin (original) data set. 1992.